

Muse: Text-To-Image Generation via Masked Generative Transformers

Year: 2023 | Citations: 677 | Authors: Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama

Abstract

We present Muse, a text-to-image Transformer model that achieves state-of-the-art image generation performance while being significantly more efficient than diffusion or autoregressive models. Muse is trained on a masked modeling task in discrete token space: given the text embedding extracted from a pre-trained large language model (LLM), Muse is trained to predict randomly masked image tokens. Compared to pixel-space diffusion models, such as Imagen and DALL-E 2, Muse is significantly more efficient due to the use of discrete tokens and requiring fewer sampling iterations; compared to autoregressive models, such as Parti, Muse is more efficient due to the use of parallel decoding. The use of a pre-trained LLM enables fine-grained language understanding, translating to high-fidelity image generation and the understanding of visual concepts such as objects, their spatial relationships, pose, cardinality etc. Our 900M parameter model achieves a new SOTA on CC3M, with an FID score of 6.06. The Muse 3B parameter model achieves an FID of 7.88 on zero-shot COCO evaluation, along with a CLIP score of 0.32. Muse also directly enables a number of image editing applications without the need to fine-tune or invert the model: inpainting, outpainting, and mask-free editing. More results are available at <https://muse-model.github.io>